What is claimed is:

1. A document descriptor extraction method comprising the steps of:

generalizing input sequences associated with a document to develop general sequences, said input sequences reflecting the structure of a document;

factoring said input sequences and said general sequences to develop factored sequences;

selecting a document descriptor from said input sequences, said general sequences, and said factored sequences using minimum descriptor length (MDL) principles.

2. The method of claim 1, wherein said selecting step comprises the steps of:

encoding said input sequences, said general sequences, and said factored sequences; and

selecting a document descriptor which encompasses all of said input sequences and exhibits a minimum MDL cost.

3. The method of claim 2, wherein said encoding step employs an algorithm which applies a set of rules comprising:

$seq(D,s) = \epsilon$ if D=s, if D does not contain metacharacters;

$seq(D_1...D_k, s_1...s_k) = seq(D_1,s_1)...seq(D_k,s_k)$;

$seq(D_1|...|D_m,s) = i\ seq(D_i,s)$;

$seq(D^*,s_1...s_k) = \{k\ seq(D,s_1)...seq(D,s_k)$ if k>0; 0 otherwise$\}$;

wherein D is a sequence of symbols, s is a sequence, and i is an index of a regular expression that the corresponding sequence s matches, wherein log m bits are needed to encode index i.

4. The method of claim 3, wherein said minimum MDL cost is determined by employing an algorithm to solve a facility location problem (FLP), said FLP modified to compute said minimum MDL cost of potential document descriptors.

5. The method of claim 4, wherein said document descriptor is a document type descriptor (DTD), and said document is an eXtensible Markup Language (XML) document.

6. The method of claim 5, wherein said minimum MDL cost comprises summing a first length of bits describing the DTD and a second length of bits for encoding the sequences.

7. A document descriptor extraction method comprising the steps of:

generalizing input sequences to develop general sequences, said input sequences reflecting the structure of data within a document;

selecting a document descriptor from said input sequences and said general sequences using minimum descriptor length (MDL) principles.

8. The method of claim 7, wherein said selecting step comprises the steps of:

encoding said input sequences and said general sequences; and

selecting a document descriptor which encompasses all of said input sequences and exhibits a minimum MDL cost.

9. The method of claim 8, wherein said encoding step employs an algorithms which applies a set of rules comprising:

$seq(D,s) = \epsilon$ if D=s, if D does not contain metacharacters;

$seq(D_1...D_k, s_1...s_k) = seq(D_1,s_1)...seq(D_k,s_k)$, if D is a concatenation of $D_1...D_k$;

$seq(D_1|...|D_m,s) = i\ seq(D_i,s)$;

$seq(D^*,s_1...s_k) = \{k\ seq(D,s_1)...seq(D,s_k)$ if k>0; 0 otherwise};

wherein D is a sequence of symbols, s is a sequence, and i is an index of a regular expression that the corresponding sequence s matches, wherein log m bits are needed to encode index i.

10. The method of claim 9, wherein said minimum MDL cost is determined by employing an algorithm to solve a facility location problem (FLP), wherein said FLP is modified to compute said minimum MDL cost of potential document descriptors.

11. The method of claim 10, wherein said document descriptor is a document type descriptor (DTD), and said document is an eXtensible Markup Language (XML) document.

12. The method of claim 11, wherein said minimum MDL cost comprises summing a first length of bits describing the DTD and a second length of bits for encoding the sequences.

13. The method of claim 7, further comprising the step of:

factoring said input sequences and said general sequences to develop factored sequences, wherein said factored sequences are available for said step of selecting;

14. A computer-readable medium encoded with a computer program for generalizing input sequences to develop general sequences, said computer program comprising:

a discover OR patterns procedure;

a discover sequence patterns procedure; and

a generalize procedure which calls said discover sequence patterns procedure and calls said discover OR patterns procedure, wherein said discover OR patterns procedure is nested within said discover sequence patterns procedure.

15. The computer-readable medium of claim 14, said computer program further comprising a partition procedure called by said discover OR patterns procedure.

16. A method for generalizing input sequences to develop general sequences comprising the steps of:

discovering OR patterns among said input sequences; and

discovering sequence patterns among said input sequences and OR patterns.

17. The method of claim 16, wherein said step of discovering OR patterns comprises the step of partitioning said input sequences.

18. A document descriptor extraction method comprising the steps of:

generalizing input sequences, said generalizing step comprising the steps of:

discovering OR patterns among said input sequences, and

discovering sequence patterns among said input sequences and OR patterns; and

selecting a document descriptor from said input sequences and said general sequences.

19. The method of claim 18, wherein said discovering OR patterns step comprises the step of partitioning said input sequences.

20. The method of claim 19, further comprising the steps of:

factoring said input sequences and said general sequences to develop factored sequences, wherein said factored sequences are available to said step of selecting.

21. The method of claim 20, wherein said step of selecting employs minimum descriptor length (MDL) principles.

22. The method of claim 21, wherein said document descriptor is a document type descriptor (DTD) and said document is an eXtensible Markup Language (XML) document.